Direct and indirect effects of continuous treatments based on generalized propensity score weighting

Yu-Chin Hsu*, Martin Huber**, Ying-Ying Lee+, Layal Pipoz++

* Academia Sinica, Institute of Economics; National Central University, Department of Finance; National Chengchi University, Department of Economics

** University of Fribourg, Department of Economics

+ University of California Irvine, Department of Economics

++ Swiss Federal Agency for Social Insurances, Mathematics Group

Abstract: This paper proposes semi- and nonparametric methods for disentangling the total causal effect of a continuous treatment on an outcome variable into its natural direct effect and the indirect effect that operates through one or several intermediate variables called mediators jointly. Our approach is based on weighting observations by the inverse of two versions of the generalized propensity score (GPS), namely the conditional density of treatment either given observed covariates or given covariates and the mediator. Our effect estimators are shown to be asymptotically normal when the GPS is estimated by either a parametric or a nonparametric kernel-based method. We also provide a simulation study and an empirical illustration based on the Job Corps experimental study.

Keywords: Mediation, direct and indirect effects, continuous treatment, weighting, generalized propensity score.

JEL classification: C21.

Correspondence: Yu-Chin Hsu, Academia Sinica, Institute of Economics, 128 Academia Road, Section 2 Nankang, Taipei, 115 Taiwan; ychsu@econ.sinica.edu.tw. Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch. Ying-Ying Lee, University of California Irvine, 3151 Social Science Plaza, Irvine, CA 92697-5100, USA; yingying.lee@uci.edu. Layal Pipoz, Bundesamt für Sozialversicherungen, Bereich Mathematik, Effingerstr. 20, 3003 Bern, Switzerland; layalchristine.pipoz@bsv.admin.ch.

1 Introduction

Classic treatment evaluations typically focus on assessing the total causal effect of a treatment on an outcome variable, e.g. the average treatment effect (ATE). In many evaluation problems, however, also the causal mechanisms appear interesting through which a total effect operates. When for example assessing the effect of an educational program on criminal activity, policy makers might want to learn whether the total effect is driven by the program's effect on employment chances which in turn may affect criminal behaviour, or by other features of the program such as its impact on personality traits like integrity or discipline. Understanding the causal mechanisms may be helpful for appropriately designing such educational programs, e.g. whether the focus should be on increasing employability, personality development, or both.

Causal mediation analysis aims to decompose a total treatment effect into the indirect effect operating through an intermediate variable called mediator, and the direct effect net of mediation; see for instance Robins and Greenland (1992) and Pearl (2001). A range of studies bases identification on conditional independence assumptions given observables with respect to treatment and mediator assignment in rather flexible (often nonparametric) models; see for instance Petersen, Sinisi, and van der Laan (2006), Flores and Flores-Lagunes (2009), van der Weele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2011), Tchetgen Tchetgen and Shpitser (2012), and Vansteelandt, Bekaert, and Lange (2012), among others.[1] Contributions concerned with nonparametric identification under conditional independence conventionally focus on binary treatments. Yet, there are many empirical problems in which treatment intensity is (close to) continuous, e.g. hours of participation in an educational program or the dose of a medical treatment.

This paper considers the identification and semi- as well as nonparametric estimation of natural direct and indirect effects (in the denomination of Pearl (2001))[2] when the treatment is continuous. The indirect effect might either concern a single mediator or reflect the impact operating through several mediators jointly and in the latter case, conditional independence must hold for each mediator. We propose an estimator based on weighting by the inverse of conditional treatment densities (i) given observed covariates and (ii) given covariates and the mediator(s), also known as generalized propensity scores; see Hirano and Imbens (2005) and Imai and van Dyk (2004). The generalized propensity scores are either obtained parametrically or nonparametrically by conditional kernel density estimation. We show that estimation is asymptotically normal and converges at the rate of one-dimensional nonparametric regression to the effects of interest under specific regularity conditions. We also provide a simulation study that illustrates the robustness of our method when compared to classic linear mediation analysis that relies on tight

[1] In contrast, the seminal papers in mediation analysis of Judd and Kenny (1981) and Baron and Kenny (1986) assume linear models for both the mediator and the outcome.

[2] Such effects have also been referred to as pure/total direct and indirect effects by Robins and Greenland (1992) and Robins (2003) or as net and mechanism treatment effects by Flores and Flores-Lagunes (2009).

parametric assumptions. Finally, we apply our approach to data on the Job Corps program, a U.S. educational intervention for disadvantaged youth. Specifically, we disentangle the program's negative effect on crime, measured by the number of arrests in the fourth year, into an indirect component operating through the mediator employment and a direct remainder effect covering any other causal mechanisms as for instance personality development. Our findings point to an important direct and nonlinear reduction of the number of arrests as a consequence of Job Corp under a sufficiently large treatment intensity of roughly 1000 hours or more, while indirect effects are close to zero for the investigated range of treatment intensities of up to 2000 hours.

Our paper fills an important methodological gap in the causal mediation literature with continuous treatment doses, where studies typically rely on rather strong functional form restrictions for identification. The semi- and nonparametric literature on continuous treatments under conditional independence is relatively sparse and focuses on the estimation of total (rather than direct and indirect) treatment effects: Flores (2007) proposes a nonparametric kernel regression estimator for average dose-response functions. Lee (2018) estimates the unconditional distribution of potential outcomes using the estimated generalized propensity score as generated regressors. Galvao and Wang (2015) propose a semiparametric propensity score weighting estimator. Our approach can be regarded as an extension of the semi- and nonparametric weighting approaches of Huber (2014) and Hsu, Huber, and Lai (2018) for discrete treatments to the continuous treatment case using kernel functions and the concept of the generalized propensity score.[3]

The remainder of the paper is organized as follows. Section 2 discusses the parameters of interest along with their identification based on weighting. Section 3 presents the estimation approach along with its properties. Sections 4 and 5 provide a simulation study based on the Job Corps experimental study, respectively. Section 6 concludes.

2 Identification

Our goal is to decompose the average treatment effect (ATE) of a continuous treatment variable D on an outcome variable Y into a direct effect and an indirect effect operating through the mediator M which may be a scalar or a vector and discrete and/or continuous. For a generic random variable A, let $\mathcal{A}$ denote the support of A. To define the effects of interest, we use the potential outcome framework, e.g. Rubin (1974), which has been applied in the context of mediation analysis by Rubin (2004), Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007), and Albert (2008), among others. Let $M(d), Y(d, M(d'))$ denote the potential mediator state as a function of the treatment and potential outcome as a function of the treatment and the potential

[3]The semiparametric version of the proposed estimator is provided in the 'causalweight' package by Bodory and Huber (2018) for the statistical software 'R'. Further alternatives that permit assessing direct and indirect effects of continuous treatments are the 'medflex' package by Steen, Loeys, Moerkerke, and Vansteelandt (2017), which implements imputation-based estimation of potential outcomes as suggested by Vansteelandt, Bekaert, and Lange (2012), and the regression-based 'mediation' package by Tingley, Yamamoto, Hirose, Imai, and Keele (2014).

mediator, respectively, under treatments values $d, d' \in \mathcal{D}$.

For each unit only one potential outcome and potential mediator state, respectively, are known, namely those related to the treatment value which is observed for that unit. That is, the observed mediator and outcome correspond to $M = M(D)$ and $Y = Y(D, M(D))$ under the observed treatment state D. In contrast, we cannot observe potential outcomes and mediators defined upon treatment values different to the observed one. Specifically, $Y(d, M(d'))$ is not observed for any individual if $d \neq d'$, as at least one of d, d' is necessarily different to the observed treatment. Identification of causal effects therefore requires specific assumptions. Similar to Imai, Keele, and Yamamoto (2010) (see their Assumption 1), Tchetgen Tchetgen and Shpitser (2012) and many others, we base identification on a sequential conditional independence assumption imposed on treatment and mediator assignment. However, contrary to the standard in the literature, we consider a continuous treatment rather than a binary one.

Our first assumption requires that given a vector of observed pre-treatment characteristics which we denote by X, the treatment is conditionally independent of the potential mediator states and the potential outcomes.

Assumption 2.1 (Conditional Independence of the Treatment):
$\{Y(d', m), M(d)\} \perp D | X = x$ *for all* $(d, d', m, x) \in \mathcal{D}^2 \times \mathcal{M} \times \mathcal{X}$.

Assumption 2.1 rules out unobserved confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on X. In the treatment or program evaluation literature, this is referred to as conditional independence, selection on observables, or exogeneity; see Imbens (2004). We point out that conditional independence must hold with respect to any value in the continuous support of the treatment, which thus appears stronger than for the binary treatment case.

Our second assumption imposes conditional independence of the mediator given the treatment and the covariates along with a common support restriction on the conditional density of the treatment. To this end, let $f_A(a|B = b)$ denote the conditional density of variable A at some value a given that variable B is equal to value b.

Assumption 2.2 (Conditional Independence of the Mediator):

(i) $Y(d', m) \perp M | D = d, X = x$ *for all* $(d, d', m, x) \in \mathcal{D}^2 \times \mathcal{M} \times \mathcal{X}$.

(ii) $f_D(d|M = m, X = x) > 0$ *for all* $(d, m, x) \in \mathcal{D} \times \mathcal{M} \times \mathcal{X}$.

Assumption 2.2 (i) rules out unobserved confounders jointly affecting the mediator and the outcome conditional on D and X. This is for instance violated if unobserved post-treatment variables influence M and Y, and are not fully determined by X and/or D. When M is multidimensional, Assumption 2.2 (i) needs to hold for each element in M, such that its strength increases in the number of mediators. Assumption 2.2 (ii) is a common support restriction. It says that the conditional density (or generalized propensity score) to receive any treatment d in

the support of D given M, X is larger than zero. This also implies that $f_D(d|X = x) > 0$ and $f_M(m|D = d, X = x) > 0$ by Bayes' theorem. Intuitively, it is required that individuals (a) with comparable values in M and X exist across all possible treatment doses and (b) with comparable values in D and X exist across all possible mediator values. We note that this assumption could be relaxed if only a subset of treatment values d was to be considered in the analysis.

Huber (2014) shows the identification of the mean potential outcomes $\mu(d, d) = E[Y(d, M(d))]$ and $\mu(d, d') = E[Y(d, M(d'))]$ with $d \neq d'$ using weighting by the inverse of specific propensity scores when Assumptions 2.1 and 2.2 are phrased in a binary context. Specifically,

$$\mu(d,d) = E\left[\frac{Y \cdot 1(D=d)}{\Pr(D=d|X)}\right], \tag{2.1}$$

$$\mu(d,d') = E\left[\frac{Y \cdot 1(D=d)}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=d'|M,X)}{\Pr(D=d|X)}\right], \tag{2.2}$$

$1(\cdot)$ denoting the indicator function. Also, $\Pr(D = d|X) = E[1(D = d)|X]$ and $\Pr(D = d|M, X) = E[1(D = d)|M, X]$ are the conditional expectations of the weights, $1(D = d)$, that correspond to the treatment propensity scores. In the binary treatment case, (2.1) and (2.2) therefore correspond to Equations (4) and (5) in Huber (2014).

Closely related identification results can be established for the case of a continuous treatment when appropriately adapting the weighting expressions; see also the discussion in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012) and Flores (2007). To this end, denote by $\omega(D; d, h)$ a weighting function that depends on the distance between D and the reference value d as well as a non-negative tuning parameter h. The closer the tuning parameter h is to zero, the less weight is given to larger discrepancies between D and d. This modification of the weighting function is required as truly continuous treatments do not have mass points. The probability of a specific value d is therefore equal to zero, which excludes the use of indicator functions. For example, as in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), we define the weighting function to be a kernel function: $\omega(D; d, h) \equiv K\left((D - d)/h\right)/h$, where K is a symmetric second order kernel function assigning more weight to observations closer to d and h is a bandwidth. Under the assumption that $f_D(d|M, X)$ and $E[Y|D = d, M, X]$ are continuous in d, the parameters of interest are identified in analogy to Equations (2.1) and (2.2) when letting h go to zero:

$$\mu(d,d) = \lim_{h\to 0} E\left[\frac{Y \cdot \omega(D;d,h)}{f_D(d|X)}\right], \tag{2.3}$$

$$\mu(d,d') = \lim_{h\to 0} E\left[\frac{Y \cdot \omega(D;d,h)}{f_D(d|M,X)} \cdot \frac{f_D(d'|M,X)}{f_D(d'|X)}\right]. \tag{2.4}$$

We note that $f_D(d|X)$ and $f_D(d|M, X)$ are the generalized propensity scores that correspond to $\lim_{h\to 0} E[\omega(D; d', h)|X]$ and $\lim_{h\to 0} E[\omega(D; d', h)|M, X]$, respectively.

The identification of the means of the potential outcomes implies the identification of the direct and indirect effects. The natural direct effect is obtained by assessing the difference in potential outcomes under two distinct treatment values, say d and d', when keeping the mediator fixed at its potential value under either d or d':

$$\theta_{d,d'}(d') = \mu(d,d') - \mu(d',d'), \quad \theta_{d,d'}(d) = \mu(d,d) - \mu(d',d), \quad \text{for } d \neq d'. \tag{2.5}$$

Equivalently, the (average) indirect effects is defined as

$$\delta_{d,d'}(d) = \mu(d,d) - \mu(d,d'), \quad \delta_{d,d'}(d') = \mu(d',d) - \mu(d',d'), \quad \text{for } d \neq d'. \tag{2.6}$$

We note that either $\theta_{d,d'}(d')$ and $\delta_{d,d'}(d)$ or $\theta_{d,d'}(d)$ and $\delta_{d,d'}(d')$ add up to the total average causal effect based on comparing potential outcomes under values d and d'. Furthermore, direct and indirect effects are permitted to be heterogeneous in M and D, respectively, as $\theta_{d,d'}(d')$ ($\delta_{d,d'}(d)$) might differ from $\theta_{d,d'}(d)$ ($\delta_{d,d'}(d')$). This allows for interactions of D and M in the determination of outcome Y.

3 Estimation

Suppose the availability of a random sample $\{(Y_i, M_i, D_i, X_i)\}_{i=1}^n$ from the joint distribution of (Y, M, D, X) for estimating the potential outcomes as well as the direct and indirect effects. We first describe fully nonparametric estimation of direct and indirect effects based on kernel methods along with its properties. At the end of this section, we discuss semiparametric estimation based on parametric generalized propensity scores. Following standard practice, the subsequent discussion implicitly assumes that regressors have been standardized by dividing by their respective standard deviations.

For a s-dimensional vector u, let $K_h(u) = \Pi_{j=1}^s k(u_j/h)/h$ be a product kernel with a generic kernel function k and bandwidth h. Let $K_{1,h_1}(u) = \Pi_{j=1}^s k_1(u_j/h_1)/h_1$ and h_1 denote the kernel function and bandwidth, respectively, for the estimation of the generalized propensity scores, and K_{2,h_2} and h_2 be the respective parameters for estimating the mean potential outcomes (based on conditioning only on D). In the first step, the generalized propensity scores, i.e., the conditional densities of D given X or M, X, are obtained by

$$\begin{aligned} \hat{f}_D(d|X_i) &= \frac{\sum_{j=1}^n K_{1,h_1}(X_j - X_i, D_j - d)}{\sum_{j=1}^n K_{1,h_1}(X_j - X_j)} \text{ and} \\ \hat{f}_D(d|M_i, X_i) &= \frac{\sum_{j=1}^n K_{1,h_1}(M_j - M_i, X_j - X_i, D_j - d)}{\sum_{j=1}^n K_{1,h_1}(M_j - M_i, X_j - X_j)}, \end{aligned} \tag{3.1}$$

respectively. In the second step, (2.3) and (2.4) are estimated by the respective sample analogs with normalized weights, which we denote by $\hat{\mu}(d,d)$ and $\hat{\mu}(d,d')$:

$$\begin{aligned} \hat{\mu}(d,d) &= \sum_{i=1}^n \frac{Y_i K_{2,h_2}(D_i - d)}{\hat{f}_D(d|X_i)} \Big/ \sum_{i=1}^n \frac{K_{2,h_2}(D_i - d)}{\hat{f}_D(d|X_i)}, \\ \hat{\mu}(d,d') &= \sum_{i=1}^n \frac{Y_i K_{2,h_2}(D_i - d)}{\hat{f}_D(d|M_i, X_i)} \cdot \frac{\hat{f}_D(d'|M_i, X_i)}{\hat{f}_D(d'|X_i)} \Big/ \sum_{i=1}^n \frac{K_{2,h_2}(D_i - d)}{\hat{f}_D(d|M_i, X_i)} \cdot \frac{\hat{f}_D(d'|M_i, X_i)}{\hat{f}_D(d'|X_i)}. \end{aligned} \tag{3.2}$$

Assumption 3.1 invokes several regularity conditions required for the consistency and asymptotic normality of the proposed estimator.

Assumption 3.1 (Regularity Conditions):

(i) *The data* $\{Y_i, M_i, D_i, X_i\}$, $i = 1, ..., n$ *are independent and identically distributed (i.i.d.).*

(ii) *The probability density function* $f_{DMX}(d, m, x)$ *is bounded away from zero and is at least* r*-order continuously differentiable with respect to* (d, m, x)*, with uniformly bounded derivatives on* $\mathcal{D} \times \mathcal{M} \times \mathcal{X}$*, a compact and convex subset of* $\mathcal{R}^{1+s_m+s_x}$*, where* s_m *and* s_x *are the dimensions of* M *and* X*, respectively.*

(iii) $E[Y|D = d, M = m, X = x]$ *is at least* r*-order continuously differentiable with respect to* (d, m, x) *on* $\mathcal{D} \times \mathcal{M} \times \mathcal{X}$ *and has uniformly bounded derivatives.*

(iv) *The symmetric kernels* k_1 *and* k_2 *are bounded differentiable, have convex bounded supports, and have order* $r_1 \geq 2$ *and* $r_2 \geq 2$*, respectively.*[4]

(v) *The bandwidths* h_1*,* h_2 *and* $h \equiv \min\{h_1, h_2\}$ *and the orders* r_1 *and* r_2 *satisfy* $h_1, h_2 \to 0$*,* $nh_1{}^{2s}h_2^2h^{-1} \to \infty$*,* $nhh_1^{4r_1}h_2^{-2} \to 0$*,* $nh_1h_2^{2r_2} = O(1)$*,* $nh_1^{2r_1+1} = O(1)$*, and* $h_1^{2r_1}h_2^{-1}h \to 0$*, as* $n \to \infty$*, where the dimension of the regressors is* $s \equiv 1 + s_m + s_x$*.*

Our estimator can be linearized to follow a U-statistic, which is well-studied in the literature. The smoothness and bandwidth conditions in Assumption 3.1 ensure that the remainder terms of the projections of the U-statistic and the bias terms are asymptotically first-order negligible. Assumption 3.1(iv) imposes standard regularity conditions for kernel functions. By Assumption 3.1(v), the second step bias is characterized by $h_2^{r_2}$, while the first step bias is dominated by terms of order $h_1^{r_1}$ and h_1^s, which involves the dimension of the regressors s. Therefore, for the first step, a higher-order kernel is required in dependence of the dimension of the regressors. For the second step, Assumption 3.1 (v) implies that one may either use the same (higher-order) kernel and bandwidth as for the first step, or alternatively a second-order kernel, requiring a smaller bandwidth $h_2 < h_1$. In the latter case, the estimation error of the first step density estimators is first-order asymptotically negligible.[5] The following theorem provides the main result of the paper, namely the asymptotic normality of our estimator.

Theorem 3.1 (Asymptotics for the Nonparametric Case)

Suppose Assumptions 2.1, 2.2 and 3.1 hold with $r \geq \max\{r_1, r_2\}$*. Then*

$$\begin{aligned}
&\sqrt{nh}\left(\hat{\mu}(d,d) - \mu(d,d)\right) \\
=&\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\left(Y_i - \mu(d,d)\right)\frac{K_{2,h_2}(D_i - d)}{f_D(d|X_i)} - \left(E[Y|D = d, X_i] - \mu(d,d)\right)\frac{K_{1,h_1}(D_i - d)}{f_D(d|X_i)} + o_p(1) \\
\xrightarrow{d}&\mathcal{N}(0, V_d),
\end{aligned}$$

where

$$V_d \equiv \begin{cases} E\left[Var[Y|D = d, X]/f_D(d|X)\right]R(k_2) \text{ if } h = h_1 = h_2 \text{ and } k_1 = k_2, \\ E\left[E\left[\left(Y - \mu(d,d)\right)^2 \middle| D = d, X\right]/f_D(d|X)\right]R(k_2) \text{ if } h = h_2 < h_1, \end{cases}$$

[4] A kernel k is of order r if $\int k(u)du = 1$, $\int u^l k(u)du = 0$ for $0 < l < r$, and $\int |u^r k(u)|du < \infty$.

[5] Furthermore, the convergence rate is slower than the rate when we use the same higher-order kernel for both steps.

in which $R(k) \equiv \int_{-\infty}^{\infty} k^2(u)du$ *and*

$$\begin{aligned}
&\sqrt{nh}\left(\hat{\mu}(d,d') - \mu(d,d')\right) \\
=&\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\Big(\left(Y_i - \mu(d,d')\right) K_{2,h_2}(D_i - d) \\
&- \left(g(d,M_i,X_i) - \mu(d,d')\right) K_{1,h_1}(D_i - d)\Big) \frac{f_D(d'|M_i,X_i)}{f_D(d|M_i,X_i) f_D(d'|X_i)} \\
&+ \left(g(d,M_i,X_i) - E\left[g(d,M,X_i)|D = d', X_i\right]\right) \frac{K_{1,h_1}(D_i - d')}{f_D(d'|X_i)} + o_p(1) \\
\xrightarrow{d}&\mathcal{N}(0, V_{dd'}),
\end{aligned}$$

where we define $g(d,M_i,X_i) \equiv E[Y|D=d,M_i,X_i]$ *and*

$$V_{dd'} \equiv \begin{cases} \Big(E\left[Var[Y|D=d,X]\frac{f_D^2(d'|M,X)}{f_D(d|M,X)f_D^2(d'|X)}\right] \\ \quad + E\left[Var[g(d,M,X)|D=d',X]/f_D(d'|X)\right]\Big) R(k_2) \text{ if } h = h_1 = h_2 \text{ and } k_1 = k_2, \\ E\left[E\left[\left(Y - \mu(d,d')\right)^2 \middle| D=d, X\right] \frac{f_D^2(d'|M,X)}{f_D(d|M,X)f_D^2(d'|X)}\right] R(k_2) \text{ if } h = h_2 < h_1. \end{cases}$$

For inference, one may use a sample analog estimator for the asymptotic variance. For example, given a uniform consistent estimator $\widehat{E}[Y|D=d,X=x]$ for $E[Y|D=d,X=x]$, a consistent for $\mathcal{V}_d$ is

$$\widehat{\mathcal{V}}_d = \frac{h}{n}\sum_{i=1}^{n}\Big(\left(Y_i - \hat{\mu}(d,d)\right) \frac{K_{2,h_2}(D_i - d)}{\hat{f}_D(d|X_i)} - \Big(\widehat{E}[Y|D=d,X_i] - \mu(d,d)\Big) \frac{K_{1,h_1}(D_i - d)}{\hat{f}_D(d|X_i)}\Big)^2.$$

This applies both when a single bandwidth is used such that $h = h_1 = h_2$ and $k_1 = k_2$ as well as when $h = h_1 < h_2$. $\widehat{\mathcal{V}}_{dd'}$ is obtained analogously.

As an alternative to basing variance estimation on the sample analogs of Theorem 3.1, one may apply bootstrap methods. Bootstrapping is known to be valid for local constant estimators; see Horowitz (2001). In the proof of Theorem 3.1, we can replace the random sample $\{(Y_i,M_i,D_i,X_i)\}_{i=1,\ldots,n}$ with the bootstrap sample $\{(Y_i^*,M_i^*,D_i^*,X_i^*)\}_{i=1,\ldots,n}$ and replace the population distribution p and E with the empirical distribution p^* and E^*.[6] Thus the bootstrap is valid in this context.

Our theory so far only considered the case in which all elements in X and M are continuous variables. We subsequently briefly discuss the inclusion of discrete variables. Consider a discrete covariate, $\widetilde{X}$, that only takes a finite number of values and enters the conditioning set in Assumptions 2.1 and 2.2 in addition to the continuously distributed X. The conditional density of $D = d$ given the covariates may be estimated by

$$\hat{f}_D(d|X_i,\widetilde{X}_i) = \frac{\sum_{j=1}^{n} 1(\widetilde{X}_j = \widetilde{X}_i) K_h(X_j - X_i) K_h(D_j - d)}{\sum_{j=1}^{n} 1(\widetilde{X}_j = \widetilde{X}_i) K_h(X_j - X_j)},$$

[6] Lemma 3.1 in Powell, Stock, and Stoker (1989) and the asymptotic linear representation for the U-statistic hold for the bootstrap estimator. The Lyapounov condition holds by the same argument.

i.e., in subcells defined upon the values of $\widetilde{X}$. Analogously, $\hat{f}_D(d|M_i, X_i, \widetilde{X}_i)$ is obtained. Replacing $\hat{f}_D(d|X_i)$ and $\hat{f}_D(d|M_i, X_i)$ in (3.2) by $\hat{f}_D(d|X_i, \widetilde{X}_i)$, and $\hat{f}_D(d|M_i, X_i, \widetilde{X}_i)$, respectively, allows estimating $\mu(d, d)$ and $\mu(d, d')$. When substituting $f_{DMX}(d, m, x)$ and $E[Y|D = d, M = m, X = x]$ by $f_{DMX\widetilde{X}}(d, m, x, \tilde{x})$ and $E[Y|D = d, M = m, X = x, \widetilde{X} = \tilde{x}]$, respectively, in Assumption 3.1, our previous asymptotic results remain valid.[7]

We conclude this section by considering semiparametric estimation of $\mu(d, d)$ and $\mu(d, d')$, in which the generalized propensity scores $f_D(d|X)$ and $f_D(d|M, X)$ are parametrically specified. To this end, we invoke the following assumption on the first step estimation of the generalized propensity scores.

Assumption 3.2 (Parametric Generalized Propensity Scores):

(i) The estimator $\hat{\gamma}_x$ of the generalized propensity score model $f_D(d|x; \gamma_x)$, $\gamma_x \in \Gamma_x \subseteq \mathcal{R}^{s_x}$, satisfies $\sup_{x\in\mathcal{X}} |f_D(d|x; \hat{\gamma}_x) - f_D(d|x; \gamma_{x0})| = O_p(n^{-1/2})$, where $\gamma_{x0} \in \Gamma_x$ such that $f_D(d|x) = f_D(d|x; \gamma_{x0})$ for all $x \in \mathcal{X}$;

(ii) The estimator $\hat{\gamma}_{mx}$ of the generalized propensity score model $f_D(d|m, x; \gamma_{mx})$, $\gamma_{mx} \in \Gamma_{mx} \subseteq \mathcal{R}^{s_{mx}}$, satisfies $\sup_{m\in\mathcal{M},x\in\mathcal{X}} |f_D(d|m, x; \hat{\gamma}_{mx}) - f_D(d|m, x; \gamma_{mx0})| = O_p(n^{-1/2})$ where $\gamma_{mx0} \in \Gamma_{mx}$, such that $f_D(d|m, x) = f_D(d|m, x; \gamma_{mx0})$ for all $m \in \mathcal{M}$ and $x \in \mathcal{X}$.

(iii) $f_D(d|x)$ and $f_D(d|m, x)$ are uniformly bounded above and bounded away from zero on $\mathcal{D} \times \mathcal{M} \times \mathcal{X}$.

A sufficient condition for Assumption 3.2 is the following. Suppose that the joint density function of D, M and X, $f_{DMX}(d, m, x)$ is uniformly bounded above and bounded away from zero and follows a parametric model such that $|f_{DMX}(d, m, x) - f_{DMX}(d, m, x; \hat{\gamma})|$ is $O_p(n^{-1/2})$ uniformly. $\hat{\gamma}$ is a root-n consistent estimator for γ_0 (typically based on maximum likelihood) with $f_{DMX}(d, m, x) = f_{DMX}(d, m, x; \gamma_0)$. Let $f_X(x)$, $f_{DX}(d, x)$, $f_{MX}(m, x)$ be the marginal density functions. Then $f_D(d|x) = f_{DX}(d, x)/f_X(x)$ and $f_D(d|m, x) = f_{DMX}(d, m, x)/f_{MX}(m, x)$, which can be consistently estimated by $f_D(d|x; \hat{\gamma}) = f_{DX}(d, x; \hat{\gamma})/f_X(x; \hat{\gamma})$ and $f_D(d|m, x; \hat{\gamma}) = f_{DMX}(d, m, x; \hat{\gamma})/f_{MX}(m, x; \hat{\gamma})$. Semiparametric estimators for $\mu(d, d)$ and $\mu(d, d')$ are given by

$$\hat{\mu}(d, d) = \sum_{i=1}^{n} \frac{Y_i K_{2,h_2}(D_i - d)}{\hat{f}_D(d|X_i; \hat{\gamma}_x)} \Big/ \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i - d)}{\hat{f}_D(d|X_i; \hat{\gamma}_x))}, \tag{3.3}$$

$$\hat{\mu}(d, d') = \sum_{i=1}^{n} \frac{Y_i K_{2,h_2}(D_i - d)}{\hat{f}_D(d|M_i, X_i; \hat{\gamma}_{mx}))} \cdot \frac{\hat{f}_D(d'|M_i, X_i; \hat{\gamma}_{mx})}{\hat{f}_D(d'|X_i; \hat{\gamma}_x)} \Big/ \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i - d)}{\hat{f}_D(d|M_i, X_i; \hat{\gamma}_{mx})} \cdot \frac{\hat{f}_D(d'|M_i, X_i; \hat{\gamma}_{mx})}{\hat{f}_D(d'|X_i; \hat{\gamma}_x)}.$$

By invoking Assumption 3.2, the asymptotic theory for these estimators simplifies considerably when compared to the nonparametric case; see Theorem 3.2 below.

[7] Note that s_x and s_m correspond to the numbers of continuous variables in X and M, respectively, i.e., without the discrete covariate $\widetilde{X}$.

Theorem 3.2 (Asymptotics for the Semiparametric Case)

Suppose Assumptions 2.1, 2.2, 3.1(i)-(iv), and 3.2 hold with $r \geq r_2$. Let the order of the kernel $r_2 = 2$. The bandwidth h_2 satisfy $h_2 \to 0$, $nh_2 \to \infty$, and $nh_2^5 \to 0$. Then

$$\begin{aligned}&\sqrt{nh_2}\left(\hat{\mu}(d,d) - \mu(d,d)\right)\\ &= \sqrt{\frac{h_2}{n}} \sum_{i=1}^{n} \left(Y_i - \mu(d,d)\right) \frac{K_{2,h_2}(D_i - d)}{f_D(d|X_i)} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V_d),\end{aligned}$$

where $V_d = E\left[E\left[\left(Y - \mu(d,d)\right)^2 \middle| D = d, X\right] / f_D(d|X)\right] R(k_2)$ *and*

$$\begin{aligned}&\sqrt{nh_2}\left(\hat{\mu}(d,d') - \mu(d,d')\right)\\ &= \sqrt{\frac{h_2}{n}} \sum_{i=1}^{n} \left(Y_i - \mu(d,d')\right) \frac{K_{2,h_2}(D_i - d) f_D(d'|M_i, X_i)}{f_D(d|M_i, X_i) f_D(d'|X_i)} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V_{dd'}),\end{aligned}$$

where $V_{dd'} = E\left[E\left[\left(Y - \mu(d,d')\right)^2 \middle| D = d, M, X\right] \frac{f_D^2(d'|M,X)}{f_D(d|M,X) f_D^2(d'|X)}\right] R(k_2)$.

The main advantage of the semiparametric approach over the fully nonparametric estimator is that it circumvents the curse of dimensionality problem when the dimensions of X and/or M are large. On the downside, misspecifications of the generalized propensity scores generally result in inconsistent estimators of potential outcomes and effects.

4 Simulation study

This section provides a simulation study to investigate the finite sample behaviour of our semi- and nonparametric methods based on the following data generating process:

$$\begin{aligned}Y &= 0.3D + 0.3M + \alpha DM + 0.3X + \beta D^3 + U,\\ M &= 0.3D + 0.3X + V, \quad D = 0.3X + W,\\ X &\sim Uniform(-1.5, 1.5),\ U, V, W \sim Uniform(-2, 2),\ \text{independently of each other.}\end{aligned}$$

Outcome Y is a function of the observed variables D, M, X and an unobserved term U. α gauges the interaction effect between D and M. $\alpha = 0$ satisfies the assumption of no interaction as discussed in Robins (2003), implying that the direct effect $\theta_{d,d'}(d) = \theta_{d,d'}(d')$ in (2.5) and the indirect effect $\delta_{d,d'}(d) = \delta_{d,d'}(d')$ in (2.6). In contrast, for $\alpha \neq 0$, direct and indirect effects are heterogeneous. β determines whether the direct effect of D on Y is linear (β=0) or nonlinear, namely cubic ($\beta \neq 0$). Mediator M is a function of D, X and the unobservable V. Note that the indirect effect is linear, as M is linear in D and Y is linear in M. Treatment D is linearly determined by X and the unobservable W. The covariate X, which confounds the treatment-outcome, treatment-mediator, and mediator-outcome relation, is continuously uniformly distributed with support ranging from -1.5 to 1.5. Finally, the unobservables follow uniform distributions with support ranging from -2 to 2. They are statistically independent of each other as well as of X.

We consider 1000 simulations and two sample sizes $n = 1000, 4000$ to investigate the performance of our nonparametric weighting approach based on (3.2). As the dimension of (D, X, M) is equal to $s = 3$ (see Section 3) in our simulation, we set the orders of the Epanechnikov kernels in (3.1) and (3.2) to $r_1 = 4$ and $r_2 = 2$, respectively. Furthermore, the bandwidth h_1 is determined by multiplying the respective standard deviations of D, X, M with $C_1 n^{-0.12}$, where $C_1 = 3.03$ is the constant term in a Silverman (1986)-type rule of thumb for fourth-order Epanechnikov kernels. Analogously, h_2 is obtained using $C_2 n^{-0.25}$, with $C_2 = 2.34$ being the constant for second-order Epanechnikov kernels. We note that these choices of r_1, r_2, h_1, h_2 satisfy the regularity conditions in Assumption 3.1 required for the satisfaction of Theorem 3.1.

Furthermore, we consider semiparametric weighting based on parametric estimation of the generalized propensity scores in (3.3). We to this end (incorrectly) assume D to be normally distributed given X or given (X, M), respectively. Bandwidth h_2 is in this case obtained using the standard rule of thumb for one dimensional kernel regression: $C_2 n^{-0.2}$, with $C_2 = 2.34$. For all kernel-based computations, we use the 'np' package by Hayfield and Racine (2008) for the statistical software 'R'. Besides estimation using bandwidths based on the rule of thumb, we consider undersmoothed versions, in which bandwidths of all kernel procedures are divided by 2.

For comparison, we in addition estimate the direct and indirect effects based on linear OLS regressions of the mediator on a constant, the treatment, and covariate and of the outcome on a constant, the treatment, the mediator, and the covariate, respectively. Concerning the definition of the direct and indirect effects, we set $d' = 0$. For d, we consider a sequence of values defined by an equidistant grid between (and including) -1.5 and 1.5 with step size 0.1 (i.e. $d \in \{-1.5, -1.4, ...1.4, 1.5\}$, however without including 0 for obvious reasons.

Table 1 reports the means of the absolute bias (abias), standard deviation (sd), and root mean squared error (RMSE) across all treatment comparisons considered for each effect under $\alpha = 0.5$ (effect hetorgeneity) and $\beta = 0$ (fully linear model). Not surprisingly, the OLS-based estimators (OLS) have the lowest standard deviations of all methods due to their parametric assumptions. On the downside, the OLS estimates of $\theta(d)$ and $\delta(d)$ are non-negligibly biased under either sample size due to the omission of the treatment-mediator interactions. In contrast, the nonparametric weighting estimator with rule of thumb bandwidths (W np) is considerably less biased. Undersmoothing (W np us) generally entails an even lower absolute bias, but as expected a higher standard deviation. A qualitatively similar pattern is observed for semiparametric weighting with a parametric first step (W p). Undersmoothing (W p us), which in the semiparametric case only concerns h_2, reduces the absolute bias and increases the standard deviation. We also note that the semi- and nonparametric versions do not uniformly dominate each other in terms of RMSE across the effects and sample sizes considered.

Table 2 gives the estimates for $\alpha = 0$ (effect homogeneity) and $\beta = 0.25$ (nonlinear direct effects). The OLS estimates of the direct effects are severely biased due to the cubic effect of D in the outcome model, while the indirect effect estimates are unbiased, as they are indeed linear. In contrast, the absolute biases of both the semi- and nonparametric weighting estimators for the direct effects are considerably smaller and decreasing in the sample size. Again, undersmoothing

Table 1: Simulations $\alpha = 0.5$, $\beta = 0$

	$\hat{\theta}(d)$			$\hat{\theta}(0)$			$\hat{\delta}(d)$			$\hat{\delta}(0)$		
	abias	sd	RMSE	abias	sd	RMSE	abias	sd	RMSE	abias	sd	RMSE
						$n = 1000$						
OLS	0.124	0.035	0.130	0.000	0.035	0.035	0.124	0.013	0.125	0.001	0.013	0.013
W np	0.020	0.057	0.062	0.062	0.056	0.086	0.077	0.010	0.077	0.039	0.007	0.040
W np us	0.016	0.101	0.103	0.044	0.100	0.113	0.048	0.035	0.060	0.023	0.024	0.034
W p	0.086	0.040	0.100	0.085	0.040	0.095	0.022	0.018	0.031	0.011	0.014	0.018
W p us	0.051	0.081	0.098	0.050	0.080	0.095	0.005	0.022	0.023	0.003	0.016	0.016
						$n = 4000$						
OLS	0.124	0.017	0.126	0.000	0.017	0.017	0.124	0.006	0.124	0.000	0.006	0.006
W np	0.016	0.038	0.044	0.054	0.037	0.069	0.065	0.008	0.065	0.034	0.005	0.034
W np us	0.021	0.063	0.067	0.043	0.062	0.079	0.048	0.021	0.052	0.026	0.014	0.029
W p	0.061	0.027	0.069	0.061	0.026	0.067	0.012	0.010	0.017	0.005	0.007	0.009
W p us	0.048	0.049	0.070	0.049	0.048	0.070	0.003	0.012	0.013	0.001	0.009	0.009

Note: 'abias', 'sd', and 'RMSE' report the the average absolute bias, standard deviation, and root mean squared error, respectively, of the effects across all treatment values $d \in \{-1.5, -1.4, ...1.4, 1.5\}$ and $d' = 0$. 'OLS', 'W np', 'W np us', 'W p', and 'W p us' refer to linear regression, nonparametric weighting, nonparametric weighting with undersmoothing in the kernel procedures, weighting with a parametric generalized propensity score, and weighting with a parametric generalized propensity score and undersmoothing in the kernel function, respectively.

generally entails a lower absolute bias than relying on rule of thumb bandwidths, but leads to higher standard deviations. Interestingly, the undersmoothed semiparametric version (W p us) dominates among all weighting approaches both in terms of small absolute biases and RMSEs, despite incorrectly assuming normality.

Finally, Table 3 provides the results when setting $\alpha = 0.5, \beta = 0.25$ (effect heterogeneity and nonlinear direct effects). Three out of four OLS effect estimates exhibit important biases, while both the semi- and nonparametric weighting estimators are less biased and superior to OLS in terms of average RMSEs under either sample size. All in all, the simulations demonstrate the merits of our methods in terms of robustness to deviations from specific parametric assumptions. This, however, comes at an efficiency cost which decreases in the sample size. The results suggest that our methods perform decently in sample sizes with several thousand observations (or more), which is quite common in empirical research.

Table 2: Simulations $\alpha = 0$, $\beta = 0.25$

	$\hat{\theta}(d)$			$\hat{\theta}(0)$			$\hat{\delta}(d)$			$\hat{\delta}(0)$		
	abias	sd	RMSE	abias	sd	RMSE	abias	sd	RMSE	abias	sd	RMSE
						$n = 1000$						
OLS	0.280	0.029	0.282	0.280	0.029	0.282	0.001	0.011	0.011	0.001	0.011	0.011
W np	0.099	0.055	0.117	0.097	0.055	0.115	0.035	0.009	0.036	0.038	0.008	0.039
W np us	0.043	0.096	0.106	0.041	0.097	0.105	0.021	0.025	0.033	0.023	0.024	0.034
W p	0.127	0.042	0.138	0.129	0.043	0.139	0.024	0.018	0.030	0.018	0.015	0.023
W p us	0.023	0.078	0.083	0.024	0.078	0.085	0.008	0.018	0.020	0.004	0.016	0.016
						$n = 4000$						
OLS	0.281	0.015	0.281	0.281	0.015	0.281	0.000	0.006	0.006	0.000	0.006	0.006
W np	0.064	0.036	0.074	0.061	0.036	0.072	0.031	0.006	0.031	0.034	0.005	0.034
W np us	0.035	0.059	0.069	0.033	0.059	0.068	0.024	0.014	0.028	0.026	0.014	0.029
W p	0.076	0.026	0.082	0.078	0.027	0.084	0.015	0.009	0.017	0.007	0.007	0.010
W p us	0.019	0.046	0.052	0.023	0.046	0.054	0.004	0.010	0.010	0.001	0.009	0.009

Note: 'abias', 'sd', and 'RMSE' report the the average absolute bias, standard deviation, and root mean squared error, respectively, of the effects across all treatment values $d \in \{-1.5, -1.4, ...1.4, 1.5\}$ and $d' = 0$. 'OLS', 'W np', 'W np us', 'W p', and 'W p us' refer to linear regression, nonparametric weighting, nonparametric weighting with undersmoothing in the kernel procedures, weighting with a parametric generalized propensity score, and weighting with a parametric generalized propensity score and undersmoothing in the kernel function, respectively.

5 Empirical illustration

We apply our method to the Job Corps study which was conducted in the mid-1990s to assess the publicly funded U.S. Job Corps program. The program targets individuals who are between 16 and 24 years, legally reside in the U.S., and come from low-income households. Participants received approximately 1200 hours of vocational training and education, housing, and board over an average duration of 8 months. Schochet, Burghardt, and Glazerman (2001) and Schochet, Burghardt, and McConnell (2008) discuss in detail the study design and report the average effects of program assignment on a broad range of outcomes. Their findings suggest that Job Corps increases educational attainment, reduces criminal activity, and increases employment and earnings, at least for some years after the program.

Several previous studies investigated various causal mechanisms of the Job Corps program. Flores and Flores-Lagunes (2009) find a positive direct effect of program assignment on earnings when controlling for the mediator work experience which they assume to be conditionally exogenous given observed covariates. Also Huber (2014) invokes a selection on observables assumption and estimates a positive direct health effect when controlling for the mediator employment.

Table 3: Simulations $\alpha = 0.5$, $\beta = 0.25$

	$\hat{\theta}(d)$			$\hat{\theta}(0)$			$\hat{\delta}(d)$			$\hat{\delta}(0)$		
	abias	sd	RMSE	abias	sd	RMSE	abias	sd	RMSE	abias	sd	RMSE
						$n = 1000$						
OLS	0.298	0.037	0.303	0.280	0.037	0.283	0.124	0.013	0.125	0.001	0.013	0.013
W np	0.100	0.061	0.122	0.114	0.060	0.132	0.076	0.011	0.077	0.038	0.008	0.039
W np us	0.044	0.102	0.112	0.056	0.101	0.120	0.047	0.035	0.060	0.023	0.024	0.034
W p	0.133	0.046	0.145	0.130	0.047	0.142	0.028	0.020	0.039	0.018	0.016	0.024
W p us	0.028	0.083	0.091	0.028	0.082	0.090	0.008	0.022	0.025	0.004	0.016	0.017
						$n = 4000$						
OLS	0.299	0.018	0.300	0.281	0.018	0.282	0.124	0.007	0.124	0.000	0.007	0.007
W np	0.064	0.039	0.076	0.078	0.038	0.089	0.065	0.008	0.065	0.034	0.005	0.034
W np us	0.035	0.063	0.073	0.049	0.062	0.083	0.047	0.021	0.052	0.026	0.014	0.029
W p	0.080	0.029	0.087	0.079	0.029	0.086	0.017	0.011	0.022	0.007	0.008	0.010
W p us	0.025	0.049	0.058	0.026	0.048	0.058	0.004	0.012	0.014	0.001	0.009	0.009

Note: 'abias', 'sd', and 'RMSE' report the the average absolute bias, standard deviation, and root mean squared error, respectively, of the effects across all treatment values $d \in \{-1.5, -1.4, ...1.4, 1.5\}$ and $d' = 0$. 'OLS', 'W np', 'W np us', 'W p', and 'W p us' refer to linear regression, nonparametric weighting, nonparametric weighting with undersmoothing in the kernel procedures, weighting with a parametric generalized propensity score, and weighting with a parametric generalized propensity score and undersmoothing in the kernel function, respectively.

Frölich and Huber (2017) use an IV strategy based on two instruments to disentangle the earnings effect of being enrolled in Job Corps into an indirect effect via hours worked and a direct effect (likely related to a change in human capital). The results point to the existence of an indirect rather than a direct mechanism. Using a partial identification approach allowing for mediator endogeneity, Flores and Flores-Lagunes (2010) derive bounds for direct and indirect effects of Job Corps assignment on employment and earnings mediated by the achievement of a GED, high school degree, or vocational degree.

Contrary to these previous contributions which consider binary treatment definitions, our interest lies in the effect of different doses, i.e. lengths of participation in Job Corps on an outcome variable capturing criminal behaviour, namely the number of arrests. Our treatment definition follows Flores, Flores-Lagunes, Gonzalez, and Neumann (2012) who assess the program's total effect on earnings. In contrast, our mediation analysis investigates whether the time spent in Job Corps affects the number of arrests indirectly through employment or 'directly', i.e. through any other causal mechanisms. More precisely, our treatment variable D is defined as the total hours spent either in academic or vocational classes in the 12 months following the program assignment according to the survey.The mediator M is the proportion of weeks employed in the second year,

while the outcome variable Y corresponds to the number of times the individual was arrested by the police in the fourth year after the random assignment.

We invoke sequential conditional independence of the treatment and the mediator as outlined in Section 2 based on a rich set of pre-treatment covariates X, which overlaps with the control variables used in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012). Specifically, we control for individual characteristics like age, gender, ethnicity, language competency, education, marital status, household size and income, previous receipt of social aid, and family background (e.g. parents' education), as well as health and health-related behavior at base line. Conditioning on such a rich set of socio-economic variables appears important, as identification relies on successfully controlling for all confounders jointly influencing at least two out of the three variables time in treatment, employment in the second year, and arrests in the fourth year. Furthermore, we condition on variables that are predictive for the duration in the program, namely expectations about Job Corps and interaction with the recruiters. Such factors appear important as they are likely correlated with personality traits like motivation, which may also affect the mediator and the outcome. Finally, we include pre-treatment outcome and mediator variables that reflect labor market and criminal behavior prior to Job Corps. This permits controlling for unobserved confounders that are time constant in the sense that they only affect the mediator and the outcome through their respective pre-treatment values.

We, however, acknowledge that our framework does not allow for dynamic confounding, implying that the length of treatment and/or the share of employment are affected by confounders that are themselves influenced by initial treatment decisions. This would for instance be the case if initial treatment participation affected motivation, which in turn influenced treatment duration, employment, and criminal behaviour. Even though we hope that the limited time horizon considered for the treatment (first year) and the mediator (second year) mitigates issues related to dynamic confounding, this threat to identification needs to be borne in mind when interpreting the results.

The original Job Corps data set consists of 15,386 individuals prior to program assignment, but a substantial share never enrolled in the program and dropped out of the study. We focus on the $10,775$ observations for which both the post-treatment variables M and Y are observed in the follow-up surveys after 2 and 4 years, respectively. There are cases of item non-response in various elements of X measured at the baseline survey, for which we account by the inclusion of missing dummies. Similar to Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), we restrict our evaluation sample to observations with a positive treatment intensity, i.e. $D > 0$, ultimately consisting of 4,000 individuals. Table 4 provides descriptive statistics for the pre-treatment covariates as well as the treatment, mediator, and outcome variables in our evaluation sample, along with the numbers of non-missing observations.

Individuals in our evaluation sample were on average 18.33 years old at baseline when applying for Job Corps and woman made up 44%. Half of the applicants were black, while whites and hispanics accounted for 25% and 17%, respectively. Regarding education, 18% of those with non-missing values held a high school diploma and 4% a General Education Diploma (GED). A large

Table 4: Descriptives

variable	mean	sd	min	max	non missing
female	0.44	0.50	0.00	1.00	4000
age	18.33	2.14	16.00	24.00	4000
white	0.25	0.43	0.00	1.00	4000
black	0.50	0.50	0.00	1.00	4000
hispanic	0.17	0.38	0.00	1.00	4000
years of education	10.05	1.54	0.00	20.00	3945
GED diploma	0.04	0.20	0.00	1.00	3982
high school diploma	0.18	0.39	0.00	1.00	3982
native English	0.86	0.35	0.00	1.00	3950
divorced	0.01	0.09	0.00	1.00	3953
separated	0.01	0.11	0.00	1.00	3953
cohabiting	0.03	0.18	0.00	1.00	3953
married	0.02	0.13	0.00	1.00	3953
has children	0.18	0.38	0.00	1.00	3981
ever worked	0.41	0.49	0.00	1.00	1405
average weekly gross earnings	19.41	98.66	0.00	2000.00	3999
is household head	0.11	0.31	0.00	1.00	3933
household size	3.52	2.01	0.00	15.00	3944
designated for nonresidential slot	0.17	0.38	0.00	1.00	4000
total household gross income (in categories)	3.51	2.21	1.00	7.00	2508
total personal gross income (in categories)	1.11	0.48	1.00	7.00	1774
mum's years of education	11.50	2.60	0.00	20.00	3263
dad's years of education	11.45	2.90	0.00	20.00	2506
dad did not work when 14	0.06	0.23	0.00	1.00	3575
received AFDC every month	0.80	0.40	0.00	1.00	1148
received public assistance every month	0.85	0.36	0.00	1.00	946
received food stamps	0.45	0.50	0.00	1.00	3836
welfare receipt during childhood (in categories)	2.07	1.19	1.00	4.00	3726
poor/fair general health status	0.13	0.33	0.00	1.00	3953
physical/emotional problems	0.04	0.20	0.00	1.00	3950
extent of marijuana use	2.54	1.55	0.00	4.00	1469
extent of hallucinogen use	2.76	1.73	0.00	4.00	204
ever used other illegal drugs	0.01	0.08	0.00	1.00	2628
extent of smoking	1.53	0.98	0.00	4.00	2084
extent of alcohol consumption	3.14	1.21	0.00	4.00	2306
ever arrested	0.24	0.43	0.00	1.00	3951
times in prison	0.07	0.35	0.00	5.00	3951
time by recruiter speaking of Job Corps (in categories)	2.05	0.94	1.00	4.00	3922
extent of recruiter support	1.59	1.07	1.00	5.00	3911
idea about wished training	0.85	0.35	0.00	1.00	3944
expected hourly wage after Job Corps	9.95	6.57	5.00	96.00	1799
expected improvement in maths	1.32	0.53	1.00	3.00	3916
expected improvement in reading skills	1.53	0.65	1.00	3.00	3932
expected improvement in social skills	1.48	0.68	1.00	3.00	3932
expected to be training for a job	1.04	0.23	1.00	3.00	3922
worried about Job Corps	0.37	0.48	0.00	1.00	3944
1st contact with recruiter by phone	0.41	0.49	0.00	1.00	3953
1st contact with recruiter in office	0.39	0.49	0.00	1.00	2315
expected stay in Job Corps	6.64	9.81	0.00	36.00	4000
total hrs spent in 1st year classes (D)	1194.15	964.89	0.86	5142.86	4000
Share of weeks employed in 2nd year (M)	44.05	37.84	0.00	100.00	4000
Number of arrests in year 4 (Y)	0.15	0.62	0.00	8.00	4000

share of respondents (had) received public assistance or welfare benefits, pointing to economic hardship. 24% had been arrested at least once prior to program assignment (excluding minor motor vehicles violations). Concerning treatment intensity (D), individuals spent on average 1194 hours either in academic or vocational classes in the first year after assignment. This corresponds to roughly 149 days of 8 hours. Thus, individuals with a positive treatment intensity were on average almost 30 working weeks in Job Corps in the first year. The treatment distribution is right skewed as the median is somewhat lower, amounting to 966 hours in classes. Concerning the share of weeks employed in the second year (M), the individuals were on average 44.05% in employment. Finally, the average number of arrests in the fourth year (Y) amounts to 0.15. Most individuals were never arrested, while 9% were arrested at least once.

We evaluate the direct and indirect effects for 20 different values of positive treatment intensity between 100 and 2000 hours in steps of 100 vs. a minor treatment of just 40 hours, corresponding to roughly one working week spent in class. That is, we estimate $\hat{\theta}(d)$, $\hat{\theta}(d')$, $\hat{\delta}(d)$, and $\hat{\delta}(d')$ for each of $d \in \{100, 200, ..., 1900, 2000\}$ and $d' = 40$, which appears sufficiently rich to approximate the quasi-continuous nature of our outcome variable.[8] Due to large number of covariates, the generalized propensity scores are estimated parametrically. We therefore assume that D is conditionally log normally distributed given X or (X, M), as it is common for non-negative treatments; see for instance Imai and van Dyk (2004). As for semiparametric weighting in Section 4, estimation relies on (3.3) and the rule of thumb for determining bandwidth h_2. We note that the obtained results are quite similar when assuming a conditional normal distribution of D (instead of log-normality) and/or applying undersmoothing by taking half of the rule of thumb bandwidth h_2. Inference is based on bootstrap standard errors obtained by bootstrapping the effects 999 times.

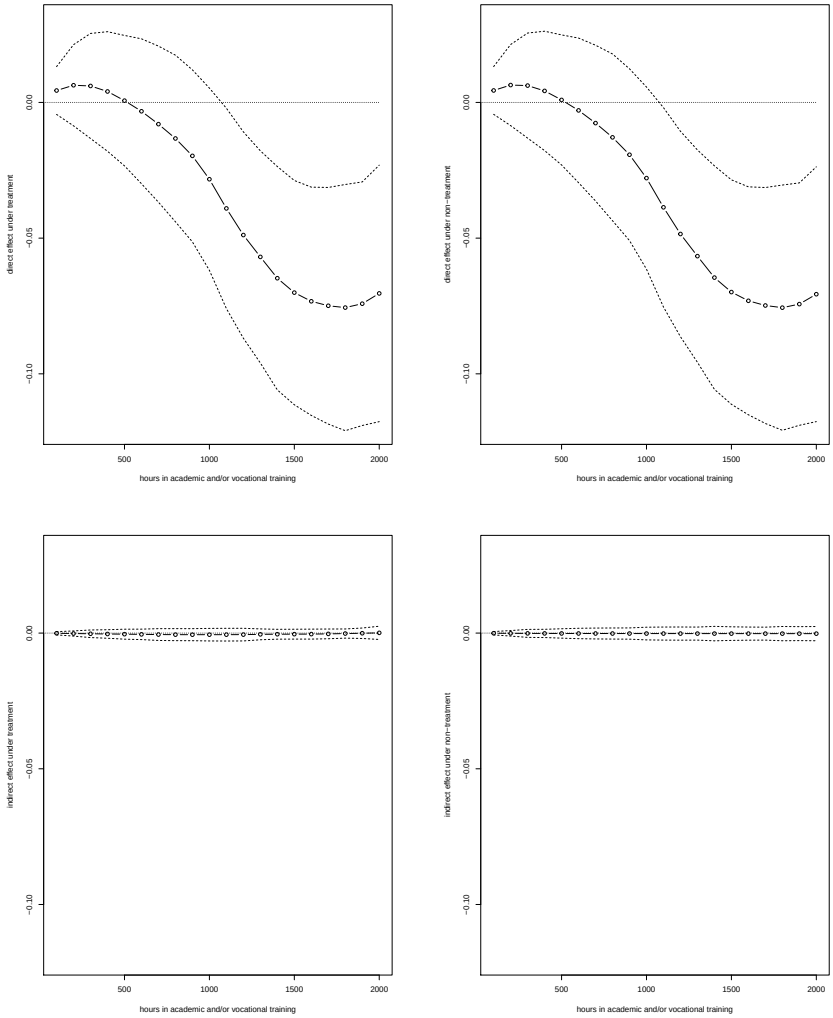
To verify whether our estimates of the generalized propensity score $f_D(d|M, X)$ successfully balance the distributions of the covariates and the mediator across treatment intensities, we conduct a test that is in the spirit of Smith and Todd (2005). Specifically, we linearly regress each of the 65 elements in X (that also include missing dummies) as well as M on the log treatment intensity, the generalized propensity score (given X and M) estimated at the sample values of D, and the score's square.[9] If (X, M) and D are not associated given the estimated propensity score such that the latter satisfies the balancing property, then the coefficient on the log treatment should be statistically insignificant in most cases. The p-values of the coefficient averages 56.7% and is only in 4 regressions (6%) smaller than 5%, such that we do not find evidence for a violation of the balancing property.

The upper panel of Figure 1 displays the direct effects under treatment (left) and non-

[8] In a robustness check, we set $d' = 0$ (no classes at all) and also include observations with zero treatment intensity in our analysis. The point estimates are similar to those presented in this section and the conclusions are the same.

[9] Using cubic or quartic polynomials of the propensity score yields similar results.

Figure 1: Direct effects (top) and indirect effects (bottom) under treatment (left) and non-treatment (right)



treatment (right). The direct effects and their marginal changes as a function of d, $\frac{\partial\theta(d)}{\partial d}$, are quite heterogeneous over the range of values d. While small treatment intensities do not appear to directly reduce the number of arrests, direct effects are statistically significantly negative at the 5% level from 1000 hours on, when the pointwise 95% confidence intervals (dashed lines) do not include zero. The effect peaks in absolute terms around 1700 hours, reducing the number of arrests by 0.07 to 0.08. In relative terms, this effect is substantial, given that the average number of arrests in the fourth year is 0.15; see Table 4.

The lower panel of Figure 1 provides the indirect effects under treatment (left) and non-treatment (right) operating through employment. All indirect effects are very small in absolute terms and never statistically different from zero at the 5% level. Summing up, our results point to an important direct, nonlinear reduction of the number of arrests in the fourth year as a consequence of Job Corp under a sufficiently large treatment intensity of roughly 1000 hours or more. In contrast, the effects of program-induced employment changes on arrests are close to zero for the investigated range of treatment intensities.

6 Conclusion

Assuming sequential conditional independence, we proposed semi- and nonparametric methods (using either parametric or nonparametric generalized propensity scores) for estimating direct and indirect effects of a continuous treatment based on inverse probability weighting and kernel methods. We demonstrated the asymptotic normality of the estimators under particular regularity conditions and investigated their finite sample behaviour in a simulation study. Finally, we applied the semiparametric method to the Job Corps program. We found this educational intervention to directly and nonlinearly decrease the number of arrests in the fourth year after assignment when controlling for employment as mediator. The semiparametric version of the proposed estimator is available in the 'causalweight' package by Bodory and Huber (2018) for the statistical software 'R'.

A Appendix

A.1 Proof of Theorem 3.1

Let the supremum norm of a function $A(z)$ be $\|A\| \equiv \sup_z |A(z)|$. Our estimator has a form of $\hat{A}/\hat{B}$. A Taylor expansion gives

$$\frac{\hat{A}}{\hat{B}} = \frac{A}{B} + \frac{\hat{A}-A}{B} - \frac{A}{B^2}(\hat{B}-B) + O_p(\|\hat{A}-A\|\|\hat{B}-B\| + \|\hat{B}-B\|^2). \tag{A.1}$$

The numerator of the estimator $\hat{\mu}(d,d)$ is

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\frac{\hat{f}_X(X_i)}{\hat{f}_{DX}(d,X_i)} \\
=&\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\left(\frac{1}{f_D(d|X_i)} + \frac{\hat{f}_X(X_i)-f_X(X_i)}{f_{DX}(d,X_i)} - \frac{\hat{f}_{DX}(d,X_i)-f_{DX}(d,X_i)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\right) \\
&+ O_p\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 K_{2,h_2}^2(D_i-d)\right) O_p\left(\|\hat{f}_{DX}-f_{DX}\|^2\right).
\end{aligned} \tag{A.2}$$

The kernel-based estimator satisfies the uniform convergence rate as in Lemma B.3 in Newey (1994),

$$\sup_{(d,m,x)\in\mathcal{D}\times\mathcal{M}\times\mathcal{X}} \left|\hat{f}_{DMX}(d,m,x) - f_{DMX}(d,m,x)\right| = O_p\Big(\Big(\frac{\log n}{n{h_1}^s}\Big)^{1/2} + {h_1}^{r_1}\Big). \tag{A.3}$$

Thus the last term in (A.2) is $O_p\left(h_2^{-1}\left((\log n/(n{h_1}^s))^{-1/2} + {h_1}^{r_1}\right)^2\right) = o_p((nh)^{-1/2})$ by Assumption 3.1(iv).

We analyze the third term in the parentheses in (A.2),

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\frac{\hat{f}_{DX}(d,X_i)-f_{DX}(d,X_i)}{f_D(d|X=X_i)f_{DX}(d,X_i)} \\
&= -\frac{1}{n}\sum_{i=1}^n \frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(\frac{1}{n}\sum_{j=1}^n K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i) - f_{DX}(d,X_i)\right) \\
&\equiv \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i} p(Z_i,Z_j) \\
&= \frac{1}{n}\sum_{i=1}^n E[p(Z_i,Z_j)|Z_i] + \frac{1}{n}\sum_{j=1}^n E[p(Z_i,Z_j)|Z_j] - E[p(Z_i,Z_j)] + Rem
\end{aligned} \tag{A.4}$$

which is a U-statistic with $Z_i \equiv (Y_i, D_i, X_i)$ and

$$p(Z_i,Z_j) \equiv -\frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\Big(K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i) - f_{DX}(d,X_i)\Big).$$

To control the remainder term Rem, we calculate

$$\begin{aligned}
&E\big[p(Z_i,Z_j)^2\big]\\
=&E\Big[\frac{Y_i^2K_{2,h_2}^2(D_i-d)}{f_D^2(d|X=X_i)f_{DX}^2(d,X_i)}E\Big[\big(K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\big)^2\Big|Z_i\Big]\Big]\\
=&O(h_2^{-1}h_1^{-s}).
\end{aligned}$$

Assumption 3.1(v) implies that $E\big[p(Z_i,Z_j)^2\big]h=O(h_2^{-1}h_1^{-s}h)=o(n)$ that further implies $Rem=o_p((nh)^{-1/2})$ by Lemma 3.1 in Powell, Stock, and Stoker (1989). The projection $E[p(Z_i,Z_j)|Z_j]$ satisfies

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n E[p(Z_i,Z_j)|Z_j]\\
=&-E\left[\frac{E\left[Y_i|D_i,X_i\right]K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(\frac{1}{n}\sum_{j=1}^n K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\right)\Big|Z_j\right]\\
=&-\frac{1}{n}\sum_{j=1}^n\frac{E\left[Y|D=d,X=X_j\right]}{f_D(d|X=X_j)}K_{1,h_1}(D_j-d)+E\left[E\left[Y|D=d,X\right]\right]+O_p(h_2{}^{r_2}+h_1{}^{r_1})\\
=&O_p((nh_1)^{-1/2}).
\end{aligned}$$

Also, the projection $E[p(Z_i,Z_j)|Z_i]$ satisfies

$$\begin{aligned}
&E[p(Z_i,Z_j)|Z_i]\\
=&-E\Big[\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\Big(K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\Big)\Big|Z_i\Big]\\
=&-\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\Big(E\Big[K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)\Big|Z_i\Big]-f_{DX}(d,X_i)\Big)\\
=&-\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\big(h_1{}^{r_1}G_i+o_p(h_1{}^{r_1})\big),
\end{aligned}$$

where $G_i\equiv\Big(\frac{\partial^{r_1}}{\partial d^{r_1}}f_{DX}(d,X_i)+\frac{\partial^{r_1}}{\partial X_i^{r_1}}f_{DX}(d,X_i)\Big)\int u^{r_1}K_1(u)du/r_1!$. The last term in (A.4) is

$$\begin{aligned}
E[p(Z_i,Z_j)]&=-E\Big[\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\Big(E\Big[K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)\Big|Z_i\Big]-f_{DX}(d,X_i)\Big)\Big]\\
&=-E\Big[\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\big(h_1{}^{r_1}G_i+o_p(h_1{}^{r_1})\big)\Big].
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n E[p(Z_i,Z_j)|Z_i]-E[p(Z_i,Z_j)]=&-\frac{1}{n}\sum_{i=1}^n\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\big(h_1{}^{r_1}G_i+o_p(h_1{}^{r_1})\big)\\
&+E\Big[\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\big(h_1{}^{r_1}G_i+o_p(h_1{}^{r_1})\big)\Big]\\
=&\,O_p(h_1{}^{r_1}/\sqrt{nh_2})\\
=&\,o_p((nh)^{-1/2}).
\end{aligned}$$

The same argument shows the second term in the parentheses in (A.2) is of smaller order. Thus we obtain the asymptotic linear representation for the numerator of $\hat{\mu}(d,d)$ in (A.2) to be

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\frac{\hat{f}_X(X_i)}{\hat{f}_{DX}(d,X_i)} - E[E[Y|D=d,X]] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(Y_i K_{2,h_2}(D_i - d) - E[Y|D=d,X=X_i]K_{1,h_1}(D_i-d)\right)/f_D(d|X_i) + o_p\left((nh)^{-1/2}\right).
\end{aligned}$$

The denominator of $\hat{\mu}(d,d)$ is equivalent to the numerator of $\hat{\mu}(d,d)$ by replacing Y_i with 1. By the same argument as above, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} K_{2,h_2}(D_i - d)\frac{\hat{f}_X(X_i)}{\hat{f}_{DX}(d,X_i)} - 1 = \frac{1}{n}\sum_{i=1}^{n}\frac{K_{2,h_2}(D_i-d) - K_{1,h_1}(D_i-d)}{f_D(d|X_i)} + o_p\left((nh)^{-1/2}\right).$$

By the Taylor expansion in (A.1), we then obtain

$$\hat{\mu}(d,d) - \mu(d,d) = \frac{1}{n}\sum_{i=1}^{n} IF_i + o_p\left((nh)^{-1/2}\right),$$

where $IF_i \equiv (Y_i - \mu(d,d))\frac{K_{2,h_2}(D_i-d)}{f_D(d|X_i)} - (E[Y|D=d,X_i] - \mu(d,d))\frac{K_{1,h_1}(D_i-d)}{f_D(d|X_i)}$. Next we show asymptotic normality by Lyapounov CLT with third absolute moment. The Lyapounov condition holds because

$$\begin{aligned}
&\left(\sum_{i=1}^{n} Var[IF_i]\right)^{-3/2}\sum_{i=1}^{n} E\left[|IF_i|^3\right] \\
=&O\left((nh^{-1})^{-3/2}\right)\sum_{i=1}^{n} E\left[|IF_i|^3\right] = O\left((nh)^{-1/2}\right) = o(1).
\end{aligned}$$

Then by the similar argument, we obtain the asymptotic variance $\lim_{n\to\infty} hVar[IF_i] = V_d$.

Now we turn to $\hat{\mu}(d,d')$. Let

$$\begin{aligned}
\hat{\Omega}_i &= \hat{\Omega}(M_i, X_i) \\
&\equiv \frac{\hat{f}_D(d'|M=M_i,X=X_i)}{\hat{f}_D(d|M=M_i,X=X_i)\hat{f}_D(d'|X=X_i)} = \frac{\hat{f}_{DMX}(d',M_i,X_i)\hat{f}_X(X_i)}{\hat{f}_{DMX}(d,M_i,X_i)\hat{f}_{DX}(d',X_i)} \\
&\equiv \frac{\hat{A}_i\hat{F}_i}{\hat{B}_i\hat{C}_i} = \Omega_i + \Omega_i\frac{\hat{A}_i - A_i}{A_i} + \Omega_i\frac{\hat{F}_i - F_i}{F_i} - \Omega_i\frac{\hat{B}_i - B_i}{B_i} - \Omega_i\frac{\hat{C}_i - C_i}{C_i} + O_p\left(\|\hat{B}_i - B_i\|^2\right).
\end{aligned}$$

We use the same argument as in the above proof for $\hat{\mu}(d,d)$. We analyze the numerator of $\hat{\mu}(d,M(d'))$, $\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i-d)\hat{\Omega}_i$. Let *s.o.* stands for smaller order terms. In the U-statistic

in (A.4), the *s.o.* are $n^{-1}\sum_{i=1}^n E[p(Z_i,Z_j)|Z_i] - E[p(Z_i,Z_j)] + Rem = o_p((nh)^{-1/2})$. Thus

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\Omega_i \frac{\hat{A}_i - A_i}{A_i} \\
=&\frac{1}{n}\sum_{j=1}^n E\left[E[Y_i|D_i,M_i,X_i]K_{2,h_2}(D_i-d)\frac{\Omega_i}{A_i}\left(K_{1,h_1}(D_j-d')K_{1,h_1}(M_j-M_i)K_{1,h_1}(X_j-X_i)-A_i\right)\middle|Z_j\right] + s.o. \\
=&\frac{1}{n}\sum_{j=1}^n E[Y_i|D_i=d,M_i=M_j,X_i=X_j]\frac{\Omega_j}{A_j}f_{DMX}(d,M_j,X_j)K_{1,h_1}(D_j-d') \\
&- E\left[E[Y_i|D_i=d,M_i,X_i]\Omega_i f_{D|MX}(d|M_i,X_i)\right] + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o. \\
=&\frac{1}{n}\sum_{j=1}^n g(d,M_j,X_j)\frac{\Omega_j B_j}{A_j}K_{1,h_1}(D_j-d') - \mu(d,d') + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.
\end{aligned}$$

By the same argument, we obtain

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\Omega_i\frac{\hat{B}_i - B_i}{B_i} \\
=&\frac{1}{n}\sum_{j=1}^n g(d,M_j,X_j)\Omega_j K_{1,h_1}(D_j-d) + \mu(d,d') + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.,
\end{aligned}$$

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\Omega_i\frac{\hat{C}_i - C_i}{C_i} \\
=&-\frac{1}{n}\sum_{j=1}^n E\left[g(d,M,X_j)|D=d',X=X_j\right]K_{1,h_1}(D_j-d')/f_D(d'|X=X_j) \\
&+ E[Y(d,M(d'))] + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.,
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\Omega_i\frac{\hat{F}_i - F_i}{F_i} \\
=&\frac{1}{n}\sum_{j=1}^n E\left[g(d,M,X_j)|D=d',X=X_j\right] - \mu(d,d') + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o = O_p(n^{-1/2}).
\end{aligned}$$

Collecting all these terms, we obtain the asymptotic linear representation for the numerator $n^{-1}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\hat{\Omega}_i$. Replacing Y_i with 1 gives the asymptotic linear representation for the denominator: $n^{-1}\sum_{i=1}^n K_{2,h_2}(D_i-d)\hat{\Omega} = n^{-1}\sum_{i=1}^n \big(K_{2,h_2}(D_i-d) - K_{1,h_1}(D_i-d)\big)\Omega_i + o_p((nh)^{-1/2})$. The Lyapounov CLT gives the asymptotic normality. □

A.2 Proof of Theorem 3.2

We first consider the $\hat{\mu}(d,d)$. Let $\Omega_i(\gamma) = 1/f_D(d|X_i)$ and $\widehat{\Omega}_i(\gamma) = 1/f_D(d|X_i;\hat{\gamma}_x)$. It is true that by mean-value expansion, $\widehat{\Omega}_i(\gamma) - \Omega_i(\gamma) = -\bar{w}_i^{-2}(f_D(d|X_i) - f_D(d|X_i;\hat{\gamma}_x))$ for some $\bar{w}_i$ between

$f_D(d|X_i)$ and $f_D(d|X_i;\hat{\gamma}_x)$. Then $\widehat{\Omega}_i(\gamma) - \Omega_i(\gamma) = O_p(n^{-1/2})$ uniformly over i. We start with the numerator of the estimator $\hat{\mu}(d,d)$. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\widehat{\Omega}_i(\gamma) \\
=&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\Omega_i(\gamma) + \frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)(\widehat{\Omega}_i(\gamma) - \Omega_i(\gamma)) \\
=&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\Omega_i(\gamma) + O_p((nh_2)^{-1/2})O_p(n^{-1/2}) \\
=&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\Omega_i(\gamma) + o_p(1),
\end{aligned}$$

where the second equality holds by a similar argument for Theorem 2 of Abrevaya, Hsu, and Lieli (2015). The derivation for the denominator follows the same arguments. By the Taylor expansion (A.1) and $E[\Omega|D = d]f_D(d) = 1$,

$$\hat{\mu}(d,d) - \mu(d,d) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \mu(d,d)}{f_D(d|X_i)}\right) K_{2,h_2}(D_i - d) + O_p\left((nh_2)^{-1}\right).$$

The asymptotic normality is shown by Lyapounov CLT with third absolute moment as the arguments in the proof of Theorem 3.1. The arguments for $\hat{\mu}(d,d')$ case are similar. □

References

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015): "Estimating Conditional Average Treatment Effects," *Journal of Business and Economic Statistics*, 33, 485–505.

Albert, J. M. (2008): "Mediation analysis via potential outcomes models," *Statistics in Medicine*, 27, 1282–1304.

Albert, J. M., and S. Nelson (2011): "Generalized causal mediation analysis," *Biometrics*, 67, 1028–1038.

Baron, R. M., and D. A. Kenny (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

Bodory, H., and M. Huber (2018): "The causalweight package for causal inference in R," *SES Working Paper 493, University of Fribourg*.

Flores, C. A. (2007): "Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment," Working paper, university of california, berkeley.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA Discussion Paper No. 4237*.

——— (2010): "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida*.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012): "Estimating the effects of length of exposure to instruction in a training program: the case of job corps," *The Review of Economics and Statistics*, 94, 153–171.

Frölich, M., and M. Huber (2017): "Direct and indirect treatment effects - causal chains and mediation analysis with instrumental variables," *Journal of Royal Statistical Society Series B*, 79, 1645–1666.

Galvao, A. F., and L. Wang (2015): "Uniformly Semiparametric Efficient Estimation of Treatment Effects with a Continuous Treatment," *Journal of the American Statistical Association*, 110, 1528–1542.

Hayfield, T., and J. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Hirano, K., and G. W. Imbens (2005): *The Propensity Score with Continuous Treatments*chap. 7, pp. 73–84. Wiley-Blackwell.

Hong, G. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects," in *JSM Proceedings, Biometrics Section*, pp. 2401–2415. American Statistical Association, Alexandria, VA.

Horowitz, J. L. (2001): "Chapter 52 - The Bootstrap," vol. 5 of *Handbook of Econometrics*, pp. 3159–3228. Elsevier.

Hsu, Y.-C., M. Huber, and T.-C. Lai (2018): "Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting," *forthcoming in the Journal of Econometric Methods*.

Huber, M. (2014): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., and D. A. van Dyk (2004): "Causal Inference With General Treatment Regimes," *Journal of the American Statistical Association*, 99, 854–866.

Imai, K., and T. Yamamoto (2011): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *unpulished manuscript*.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Lee, Y.-Y. (2018): "Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models," arxiv:1811.00157.

Newey, W. (1994): "Kernel estimation of partial means and a general variance estimator," *Econometric Theory*, 10, 1–21.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of Direct Causal Effects," *Epidemiology*, 17, 276–284.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–30.

Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (2004): "Direct and Indirect Causal Effects via Potential Outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Schochet, P. Z., J. Burghardt, and S. Glazerman (2001): "National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes," *Report (Washington, DC: Mathematica Policy Research, Inc.)*.

Schochet, P. Z., J. Burghardt, and S. McConnell (2008): "Does Job Corps Work? Impact Findings from the National Job Corps Study," *The American Economic Review*, 98, 1864–1886.

Silverman, B. (1986): *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Smith, J., and P. Todd (2005): "Rejoinder," *Journal of Econometrics*, 125, 365–375.

Steen, J., T. Loeys, B. Moerkerke, and S. Vansteelandt (2017): "Medflex: an R package for flexible mediation analysis using natural effect models," *Journal of Statistical Software*, 76.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *The Annals of Statistics*, 40, 1816–1845.

Ten Have, T. R., M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck (2007): "Causal mediation analyses with rank preserving models," *Biometrics*, 63, 926–934.

Tingley, D., T. Yamamoto, K. Hirose, K. Imai, and L. Keele (2014): "Mediation: R package for causal mediation analysis," *Journal of Statistical Software*, 59, 1–38.

van der Weele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.